3/12/15

lecture 13

~~Kleinberg~~ Clustering impossibility ~~An impossibility result for clustering~~

Bibliographic note: the material starting w/ the stopping conditions for single-link clustering all comes from Kleinberg, NIPS 2003. (see webpgs for link & full bib. info).

Announcement re: grades for A1.

In our explorations of clustering, we've taking on "optimization function"-based perspective:

~~- each clustering algo would~~

So far: ~~each clustering algo each had diff optimization fn, and produced~~ produces k clusters

~~- each produ~~

~~3 clustering ideas~~

3 ideas for producing ~~k clusters~~ clusterings, w/ associated optimization function.

~~<say, to prevent students feeling like~~

<say, don't write, the rest so that we don't spend too much time in review>.

- k-means: minimize ~~squared~~ within-cluster squared distances. can solve approximately
- single-link: maximize the smallest spacing btwn clusters can solve exactly.
- spectral clustering: find embedding that gives smallest normalized cut. ~~solve~~ approximate solution. >

~~Now, if we have too~~

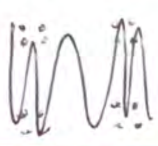
~~Now, I've only~~

This lecture, going to ~~thoroughly~~ look @ these premises more carefully:

k fixed so far, but shouldn't ~~this~~ ~~k vary~~? it ~~depend on~~ vary depending on the data?

which one best when?

ex: ~~picking k~~ ~~sps we choose~~ k = 3, ~~and our data looks like this~~:



"ought" to be 4 clusters vs. "ought" to be 3

⇒ best to allow # of clusters to vary.

Consider the following single fix to allow ~~adapting~~ adjusting the # of clusters on the fly:

" for given opt. fn, ~~try all possible # of clusters~~ for k ( ), ~~and pick the best clustering~~ get the <approximate> best clustering for each k, and then pick the best among those.

let k^* = # of clusters in that clustering

- best for ~~k-means~~ w/in cluster sum of squared distances: ~~everyone in their own cluster~~ @ $k^* = n$
- best for ~~single-link~~: ~~merging the inter-cluster spacing~~
  - ↳ maximizing the between-cluster spacing: ~~furthest 2 points in diff. clusters (doesn't matter about anyone else)~~

<what does this mean if there aren't 2 diff clusters?

Ill-defined for 1 cluster only. But o.w., it works and you just pick the best and that will be two clusters

$k^* = 2$

- best for normalized cut:

$$\frac{(\#\text{ of edges } \sout{\text{cut}})}{\sum_j (\#\text{ edges inside } C_j)}$$

: ~~one big cluster~~. $k^* = 1$.

- maybe best to skip all this: it would require us to write down all the objective functions anyway, ~~and the only "problem" that results from single-link clustering is the method being ill-defined @ k=1~~ (w.r.t. objective fn).
- could ask as matching ~~clicker~~ clicker-q?

~~Let k^* be the # of clusters in the optimal clustering~~

Let $k^* = \arg\max_k$ best

Let k^* = # of clusters in the ~~optimal~~ clustering optimizing the specified function.

What is k^*?.. (Assume $n >> 2$)

Q1: for k-means' opt. fn?

- ~~if only~~

(A) $k^* = 1$

(B) $k^* = 2$

~~(C) $k^* = 3$~~

~~(D)~~ $k^* = n - 1$

(E) $k^* = n$

(F) I don't know.

Q2: for normalized cut?

Q3: for single-link? (~~we do not allow~~ ignore $k^* = 1$).

Need extra criterion for sensible output.

So quickly! but good for them to remember these fns.

So, what else might we try? Need sthg about the data points to guide us.
~~To develop intuition, let's see how single-link as a concrete example, and consider how we might get it to stop~~

clicker q. 4: (diff. possible stopping conditions for single-link)
which of the following adaptations of single-link clustering is the "best" way to allow it to ~~create diff~~ produce different #s of clusters?

(A) Instead of stopping cluster merging when there are k connected components, stop when there are no more diff-cluster pairs of points with distance $\leq \theta$; ← let θ be a fixed positive #, and

~~(A) ... for fixed threshold θ, stop when there are no more diff-cluster pairs~~ with distance $\leq \theta$;

(A) ... for fixed threshold θ, stop when there are no more diff-cluster pairs with distance $\geq \theta$

(B) ~~for fixed multiplier of, stop when there are~~ that have distance $\leq \theta \cdot \max_{i,s} d(x_i, x_s)$ ~~(max diff between cluster~~

(C) both would be good ~~but~~ ~~are good, except go have the inequality in (B) going the wrong way~~. if you changed the "$\geq$" in (B) to "$\leq$"

(D) both would be good if you changed

(c) ~~Exactly one of the "$\leq$"~~ Neither are good, but exactly one would be good if you changed a "$\leq$" to a "$\geq$" but both would be if you changed.

(D) Neithr are good, but both would be if you changed both "$\leq$" to "$\geq$"

(E) what's single-link clustering?

(note: this allows extension to $k^* = 1$)

~~note that we lose our guarantees~~

~~Other examples:~~
~~For other~~
~~could try to do sthg similar~~

skip
- Other alternatives:
  - add penalty term to avoid overfitting (k-means)
  - ~~regularization~~
  - eigengap in the spectrum (list of eigenvalues)

~~All these~~ NOW

~~All these questions about trying to~~

All these questions ~~were motivated~~ about how one might go about varying the # of clusters were motivated by the principle that one really can't a priori know how many clusters ~~to~~ one wants.

~~We might encode this~~ enshrine this principle by giving it a name:

~~Formalizing the~~ the idea that ~~# of clusters need not be fixed~~ any # of

Let $X = \{x_1, \ldots, x_n\}$ our points.

$d: \{1, \ldots, n\} \times X \times X \to \mathbb{R}$ a distance function.

And this principle operated "above" the level of specific optimization fns.

~~Idea:~~ ~~are these principle~~

Principle: whatever optimization function ~~of~~ alg we choose, it ~~should be able~~ it should be able to produce any possible clustering:

We can make this principle even more general, and say that not only should the number of clusters be allowed to vary over all possible #s, but that the clusterings themselves should be allowed to vary over all possible partitions.

Our work suggests:

A meta-principle (~~rather than an opt. fn~~):

~~Exercise~~

for convenience let's number points, and say $X = \{1, 2, \ldots, n\}$.

For any partition Γ of X, we can set up points where that's the 'natural' clustering

ex: $\Gamma = \{1,2,3\}, \{4,5,6\}$

ex: $\Gamma = \{1,2\}, \{3,4,5\}, \{6\}$

distances

$d: X \times X \to \mathbb{R}^{\geq 0}$

So, define a clustering fn as: f: input = ~~n~~ distance fn ~~d(i,j)~~ on points
output: a partition of X.

Richness ~~property~~ property: $Range(f)$ = set of all partitions of X.

<whatever an alg is, it can't rule out some clustering a priori>

[Cornell's] Jon

<~~note that all our fixed-k algs violated this.~~

Kleinberg's insight: meta-analysis: level above specific opt. fns,

~~Note:~~ Note: all our fixed-k algs violated this. ~~But this~~ we talked about fixes

<sthg we'll get back to later...>

Now, while we're @ it, let's look @ other reasonable properties.

If the only input we have is the distance function onto the non-negative #'s — note there are no "units", then it shouldn't matter what the units are.

(no ~~units~~ "units")

~~Scale~~

Scale-invariance property: for any distance fn d and any $\alpha > 0$, let $d_\alpha(t,s) = \alpha d(t,s)$ ("scale it").

$f(d) = f(d_\alpha)$

~~partition you get~~ result partition ~~resulting partition~~

<if you spread all the distances by the same amt and get same result>.

<all the distances have the same relative relationships to them>

<seems like k-means, single-link, (tie!)

<drawing a picture actually makes it seem more suspect>

Consistency property: [if you shrink "w/in-cluster distances"; expand "between-cluster distances"; f gives the same partition out.]

<data is only becoming even more well-clustered>

~~- this "defn" requires introducing a specific partition.~~

Let d be any distance function:

$P = f(d)$ the output partition

and d' = any distance function where t,s are in the same cluster in P ($(t,s) \sim P$)

$\Rightarrow d'(t,s) \leq d(t,s)$.

"shrinks inter-cluster distances"

and t,s in diff clusters in P ($(t,s) \not\sim P$)

$\Rightarrow d'(t,s) \geq d(t,s)$

"expands btwn cluster distances"

~~f satisfies~~

Then, $f(d') = P$ $(= f(d))$

P is f look too similar

Q5: which does our fixed-k single-link clustering satisfy?

(0): B ; C

(A) just richness
(B) just scale-invariance
(C) just consistency
(D) exactly 2
(E) all 3.

Q6: which does our distance-0 single-link clustering satisfy?

(0): A ; C

Q7: which does ~~one~~ our fixed-k-means alg satisfy?

- maybe requires pf! (B) ; (C).
<not (A) is obvious>

OK: so, is there an alg that satisfies all 3 properties?
Well, clearly spectral clustering, right? o.w., why would we have ~~done~~ presented it last?

outline:

thm 3.1: spsc clustering fn f satisfies scale-invariance ; consistency.
Then, Range(f) does not contain any pair of partitions where one refines the other.

We ~~say that Range~~
That is, Range(f) is an anti-chain. |vocab item.

[not only is it not a chain:

$\{1,2\}\{3\},\{4,5\} \leq \{1,2,3\}\{4,5\} \leq \{1,2,3,4,5\}$

then isn't even any "links" (a pair)]

thm 3.2: for every antichain A, there is a clustering fn f s.t. Range(f) = A and f satisfies scale-invariance ; consistency.

~~Since the set of all partitions is not an anti-chain~~

~~Since~~
But, the set of all partitions is not an antichain.
⇒ can't have both richness ; the other two simultaneously.
(so, no, spectral clustering doesn't satisfy all three)

~~<that explains>~~

- a formal expression of fundamental tradeoffs in clustering.
<so instead of fighting over which opt. fn is best, realize each is giving up something>.

- let's do 3.2:
shows how we can satisfy scale-invar & consistency ... as long as we restrict the possible outputs.
uses our k-means opt-fn
~~shows~~

~~Let f in our anti-chain of partitions.~~

Let f output the partition $P \in A$ that minimizes:

$$\Phi_d(P) = \sum_{\substack{t,s \text{ in} \\ \text{same cluster in } P}} d(t,s)$$

(recall: w/in cluster scatter, for d = Euc. dist. squared)

- b/c A an anti-chain, all-singleton clusters ~~disallow~~ $\notin A$.

scale-invariance? ✓ spreading all the distances by same amt ~~means~~ would still yield same minimum.

~~$\Phi_d(P)$ has same min as~~ would just have the scaling factor here.

~~Range(f)~~ $\overset{c}{=} A$? ✓ by ~~construc~~ defn.

prove: $A \subseteq Range(f)$. to get that we can get this "pseudo-richness".

i.e: show: for any $\hat{P} \in A$, there's a distance matrix d so that $f(d) = \hat{P}$.

just make all in-cluster distances really small, ex: $d(t,s) < \frac{1}{n^3}$ if $t,s \sim \hat{P}$.

btwn- " " relatively big, ex: ~~$d(t,s) > 1$~~.

→ so, $\Phi_d(\hat{P}) = \sim n^2$ sum of $< \frac{1}{n^3}$ terms, < 1.

clicker Q8: for btwn-cluster distances, set $d(t,s)$:
(A) $\frac{1}{n}$ (B) $\frac{1}{n^2}$ (C) 1 (D) IDK.

=> for f to output something P', ~~other than~~ P, that ~~"something" must also~~ $\Phi_d(P')$ must be < 1

=> ~~can~~ P' must refine $\hat{P}$. (o.w, would have an edge

after stuff below

- need $\hat{P}$ to be ~~the minimizer of~~ $\operatorname{argmin}_{P \in A} \Phi_d(P)$ (should use the anti-chain fact

- if can show any P' w/ $\Phi_d(P') < \Phi_d(\hat{P})$, P' is a refinement of $\hat{P}$, done. (b/c then $\hat{P} \notin A$,

* idea: force any 'minimizer' to not have any btwn-cluster pairs be w/in a clust. of $\hat{P}$

=>

consistency? (let's skip).

to show: for a given d, let $\hat{P} = f(d)$.

take any d' s.t. w/in cluster for P squished, btwn-cluster for P increased.

Show: $f(d) = f(d')$

~~we know $\hat{P} = \arg\min_P \Phi_d(P)$, show $\hat{P} = \arg\min_P \Phi_{d'}(P)$~~

(A) for any other P', $\Phi_{d'}(P') \geq \Phi_{d'}(\hat{P})$

$\Phi_{d'}(P')$: at least $\Phi_d(P')$; $\Phi_{d'}(\hat{P})$: smallest for d.

we know $\Phi_d(P') \geq \Phi_d(\hat{P})$ b/c of defn of f.

want: $\Phi_{d'}(P') \geq \Phi_{d'}(\hat{P})$.

so take $\Phi_{d'}(P') - \Phi_{d'}(\hat{P})$ vs. $\Phi_d(P') - \Phi_d(\hat{P})$.

$$= \sum_{\substack{\text{w/in cluster} \\ \text{pairs for } P' \\ \text{i.e. } s,t \sim P'}} \left[d(s,t) - d'(s,t) \right]$$

$$\leq \sum_{\substack{s,t \sim P' \\ \text{and} \\ s,t \not\sim \hat{P}}} \left[d(s,t) - d'(s,t) \right]$$

~~added this here:~~

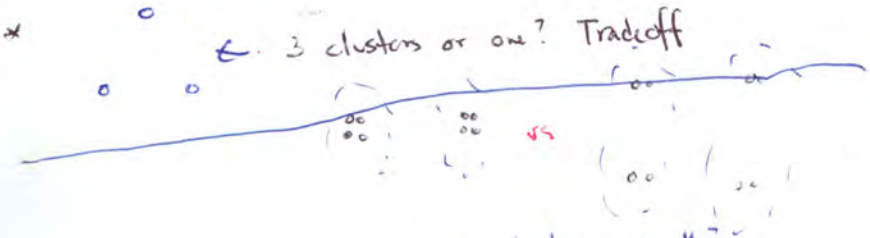
for edges ~~within~~ not btwn in $\hat{P}$ but in P within, $d'(s,t) > d(s,t)$ b/c of d' being a P-transform

$$\leq \sum_{s,t \sim \hat{P}} \left[d(s,t) - d'(s,t) \right]$$

added more edges that are $\sim$ in $\hat{P}$, so non-neg. quant is added.

$$= \Phi_d(\hat{P}) - \Phi_{d'}(\hat{P}), \text{ as desired.}$$

~~Not~~ ← 3 clusters or one? Tradeoff



vs.

2 clusters vs. 4;

The point is not that the properties are "bullet pf". The point is to be able to formalize the tradeoffs.